The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.
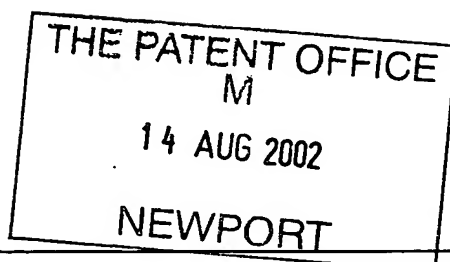
Signed

Dated   14 November 2002

An Executive Agency of the Department of Trade and Industry

The Patent Office

Patents Act 1977

Rule 16

Request for grant of a patent

THE PATENT OFFICE
M

1 4 AUG 2002

NEWPORT

The Patent Office
Concept House
Cardiff Road
Newport
South Wales NP10 8QQ

1.	Your reference	GB920020033GB1

2.	Patent application number *(The Patent Office will fill in this part)*	**0218891.0**	I 4 AUG 2002

3.	Full name, address and postcode of the or of each applicant *(underline all surnames)*	INTERNATIONAL BUSINESS MACHINES CORPORATION Armonk New York 10504 United States of America
	Patents ADP number *(if you know it)* S19637001	
	If the applicant is a corporate body, give the country/state of its incorporation	State of New York United States of America

4.	Title of the invention	METHOD FOR DATA RETENTION IN A DATA CACHE AND DATA STORAGE SYSTEM

5.	Name of your agent *(if you have one)*	R J Burt
	"Address for Service" in the United Kingdom to which all correspondance should be sent *(including the postcode)*	IBM United Kingdom Limited Intellectual Property Department Hursley Park Winchester Hampshire S021 2JN
	Patents ADP number *(if you know it)* 790392001	

		Country	Priority App No *(if you know it)*	Date of filing *(day/month/year)*
6.	If you are declaring priority from one or more earlier patent applications, give the country and the date of filing of the or of each of these earlier applications and *(if you know it)* the or each application number			

		No of earlier application	Date of filing *(day/month/year)*
7.	If this application is divided or otherwise derived from an earlier UK application, give the number and the filing date or the earlier application		

THE PATENT OFFICE
M
14 AUG 2002
NEWPORT

The
Patent
Office
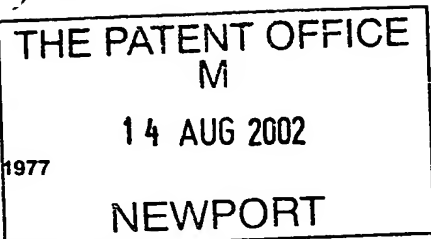
Patents Act 1977
Rule 15

Statement of inventorship and of right to grant of a patent
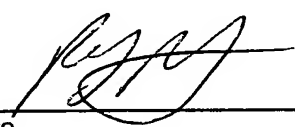
The Patent Office
Concept House
Cardiff Road
Newport
South Wales NP10 8QQ

1.	Your reference	GB920020033GB1
2.	Patent application number *(if you know it)*	**0218891.0** 14 AUG 2002
3.	Full name of the or of each applicant	INTERNATIONAL BUSINESS MACHINES CORPORATION
4.	Title of invention	METHOD FOR DATA RETENTION IN A DATA CACHE AND DATA STORAGE SYSTEM
5.	State how the applicant(s) derived the right from the inventor(s) to be granted a patent	By employment and agreement
6.	How many, if any, additional Patents Forms 7/77 are attached to this form?	1

7.

I/We believe that the person(s) named over the page (and on any extra copies of this form) is/are the inventor(s) of the invention which the above patent application relates to.

Signature

13 August 2002
Date

8.	Name and daytime telephone number of person to contact in the United Kingdom	R J Burt Tel: 01962 816646

Patents Form 7/77

Enter the full names, addresses and postcodes of the inventors in the boxes and underline the surnames

Paul <u>ASHMORE</u>
(UK Resident)
c/o IBM United Kingdom Limited
Intellectual Property Law
Hursley Park
Winchester
Hampshire SO21 2JN
England

Patents ADP number *(if known)*

Michael Huw <u>FRANCIS</u>
(UK Resident)
c/o IBM United Kingdom Limited
Intellectual Property Law
Hursley Park
Winchester
Hampshire SO21 2JN
England

Patents ADP number *(if known)*

If there are more than three inventors, please write their names and addresses on the back of another Patents Form 7/77 and attach it to this form

Robert Bruce <u>NICHOLSON</u>
(UK Resident)
c/o IBM United Kingdom Limited
Intellectual Property Law
Hursley Park
Winchester
Hampshire SO21 2JN
England

Patents ADP number *(if known)*

REMINDER

Hav you signed the form?

METHOD FOR DATA RETENTION IN A DATA CACHE

AND DATA STORAGE SYSTEM

This invention relates to data storage systems. In particular, this
invention relates to a method and system for data retention in a data
cache.

In existing, well-known write caching systems, data is transferred
from a host into a cache on a storage controller. The data is retained
temporarily in the cache until it is subsequently written ("destaged") to a
disk drive or RAID array.

In order to select the region of data to destage next, the controller
firmware uses an LRU (Least Recently Used) algorithm. The use of an LRU
algorithm increases the probability of the following advantageous events
happening to the data in the cache.

1.   Data in the cache may be overwritten with updated data before
being destaged, so that write operations from the host result in only one
destage operation to the disk, thereby reducing disk utilisation.

2.   Data in the cache may be combined with logically-adjacent data
(coalesced) to form a complete stride for destaging to a RAID 5 array,
thereby avoiding the read-modify-write penalty typically encountered when
writing to a RAID 5 array.

3.   An attempt by the host to read data which it has recently written may
be serviced from the cache without the overhead of retrieving the required
data from the disk. This improves the read response time.

Data in the cache must be protected against loss during unplanned
events (e.g. resets or power outages). This is typically achieved by
including battery backed memory or UPS (uninterruptible power supply) to
allow the data to be retained during such events.

However, the provision of such backup power is difficult and
expensive so a design decision is often taken such that the controller may
not have sufficient power available to retain the contents of all of its
cache memory. Consequently, the controller has areas of cache memory which
cannot be used for write caching (since the data stored therein would be
vulnerable to loss).

data may be added to the head of the second list when the data is destaged.
If the data was not read when referenced in the first list, the data may be
either maintained in its position in the second list or discarded.

5   The flag may include a timestamp each time the data is read and the
timestamp may be used to prioritise the position of the data reference in
the second list.

Data may be partly dirty and partly clean and may be referenced in
10   both the first and second lists.

According to a second aspect of the present invention there is
provided a data storage system comprising: a storage controller including a
cache; a data storage means; and the cache has a first least recently used
15   list for referencing dirty data which is stored in the cache, and a second
least recently used list for referencing clean data; wherein dirty data is
destaged from the cache when it reaches the tail of the first least
recently used list and clean data is purged from the cache when it reaches
the tail of the second least recently used list.
20
Dirty data which is destaged to a data storage means may have a copy
of the data retained in the cache as clean data which is deleted from the
first list and added to the second list.

25   A read command which is a cache miss may fetch data from the data
storage means and the data may be retained in the cache with a reference in
the second list.

A flag may be provided with each data reference in the first list
30   indicating whether or not the data has been read whilst on the first list.
If the data was read when referenced in the first list, the data may be
added to the head of the second list when the data is destaged. If the
data was not read when referenced in the first list, the data may be either
maintained in its position in the second list or discarded.
35
The flag may include a timestamp each time the data is read and the
timestamp may be used to prioritise the position of the data reference in
the second list.

40   Data may be partly dirty and partly clean and may be referenced in
both the first and second lists.

According to a third aspect of the present invention there is
provided a computer program product stored on a computer readable storage

may have subsets referred to as pages. In an example implementation, a
page is 4k bytes giving 16 pages in a track. Each of the pages in a track
may be dirty, clean or absent. In practice, there may also be subsets of
pages.

5

   The first list 104 is for dirty data which is data that has been
received from the host 101. The first list 104 is referred to as the LRW
(Least Recently Written) list. The second list 105 is for clean data which
is data which has been destaged to the data storage means 106 and a copy is
10   retained in the cache 103. The second list 105 is referred to as the LRR
(Least Recently Read) list.

   Referring to Figure 2, a detail of Figure 1 is provided showing the
cache 103 with the LRW list 104 and the LRR list 105. A data region in the
15   cache 103 will always be on at least one list 104, 105 and may be on both
lists.

   When the dirty data is initially stored 200 in the cache 103, a
corresponding entry 201 is created for it on the dirty LRW list 104. When
20   the data is destaged and marked clean, it is deleted from the LRW list 104
and added 202 to the LRR list 105.

   Additionally, a data region may be partly dirty and partly clean. As
described above, a data region in the form of a track may have some dirty
25   pages and some clean pages. In this case the track would be on both lists
104, 105, since it must be possible to find it both when searching for a
destage candidate and when searching for a purge candidate. Individual
pages can be destaged or purged, rather than doing this at track level.

30   There is also another route onto the LRR list 105. In a general
read/write cache 103, there are read commands from the host 101 which are
cache misses. In this case, data is fetched from the data storage means
106 and may be retained in the cache 103 to satisfy further read commands
from the host 101. A corresponding entry 203 is made for the data on the
35   LRR list 105.

   This is particularly beneficial in an environment where the storage
controller 102 may be accessed from multiple hosts, since multiple hosts
often utilise some regions of the disks for storing shared data and
40   consequently multiple hosts may read the same disk region frequently.

   There is a problem of how to assign suitable priority to data which
was dirty but has been destaged so is now marked as clean. This data
region needs to be deleted from the LRW list and, potentially, added to the

has been read whilst dirty, the data descriptor is sent 308 to the head of
the LRR list. If the data has not been read whilst dirty, the data is
discarded.

5   The following is a detailed description of the described method. The
following should be noted.
   · Virtual Track (VT) is the jargon used for a data region in the cache,
     which contains some dirty data, some clean data or both.
   · Cache directory (CD) is the jargon used for the overall directory of
10   cache elements.
   · To be considered for a read or write hit, or for destaging or
     purging, a VT must be in the CD.

Two queues are maintained:
15

LRW queue of VTs with ANY pages containing some dirty data.
LRR queue of VTs with ANY pages containing no dirty data.

General Rules:
20

   · VTs get added/moved to the head of the LRW queue whenever they are
     populated with one or more dirty sectors.

   · VTs get added/moved to the head of the LRR queue whenever they are
25   read and contain a clean page.

   · VTs which get read have their "read" flag set.

   · When a VT which is not already on the LRR queue is destaged and
30   marked clean, it is added to the head of the LRR queue if the "read"
     flag is set. Otherwise it is deleted.

Rules in detail:

35 Dirty VT inserted into CD:
        The VT is added to the head of the LRW queue.

Clean VT inserted into CD:
        The VT is added to the head of the LRR queue.
40
Dirty data merged into VT in LRW queue:
        The VT is moved to the head of the LRW queue.

Dirty data merged into VT in LRR queue:

The described method particularly improves write performance for RAID 5 storage arrays by permitting data coalescing into full-stride writes.

The described technology could be used in disk drives, disk controllers/adapters and file servers.

Modifications and improvements may be made to the foregoing without departing from the scope of the present invention.

8.   A method as claimed in any one of the preceding claims, wherein data is partly dirty and partly clean and is referenced in both the first and second lists (104, 105).

5   9.   A data storage system comprising:

       a storage controller (102) including a cache (103);

       a data storage means (106); and

10

       the cache (103) has a first least recently used list (104) for referencing dirty data which is stored in the cache (103), and a second least recently used list (105) for referencing clean data;

15   wherein dirty data is destaged from the cache (103) when it reaches the tail of the first least recently used list (104) and clean data is purged from the cache (103) when it reaches the tail of the second least recently used list (105).

20   10.   A data storage system as claimed in claim 9, wherein dirty data which is destaged to a data storage means (106) and a copy of the data is retained in the cache (103) as clean data is deleted from the first list (104) and added to the second list (105).

25   11.   A data storage system as claimed in claim 9 or claim 10, wherein a read command which is a cache miss fetches data from the data storage means (106) and the data is retained in the cache (103) with a reference in the second list (105).

30   12.   A data storage system as claimed in any one of claims 9 to 11, wherein a flag is provided with each data reference in the first list (104) indicating whether or not the data has been read whilst on the first list (104).

35   13.   A data storage system as claimed in any one of claims 9 to 12, wherein, if the data was read when referenced in the first list (104), the data is added to the head of the second list (105) when the data is destaged.

40   14.   A data storage system as claimed in any one of claims 9 to 13, wherein, if the data was not read when referenced in the first list (104), the data is either maintained in its position in the second list (105) or discarded.

ABSTRACT

METHOD FOR DATA RETENTION IN A DATA CACHE

AND DATA STORAGE SYSTEM

5

A method for data retention in a data cache and a data storage system are provided. The data storage system (100) includes a storage controller (102) with a cache (103) and a data storage means (106). The cache (103) has a first least recently used list (104) for referencing dirty data which is stored in the cache (103), and a second least recently used list (105) for clean data in the cache (103). Dirty data is destaged from the cache (103) when it reaches the tail of the first least recently used list (104) and clean data is purged from the cache (103) when it reaches the tail of the second least recently used list (105).
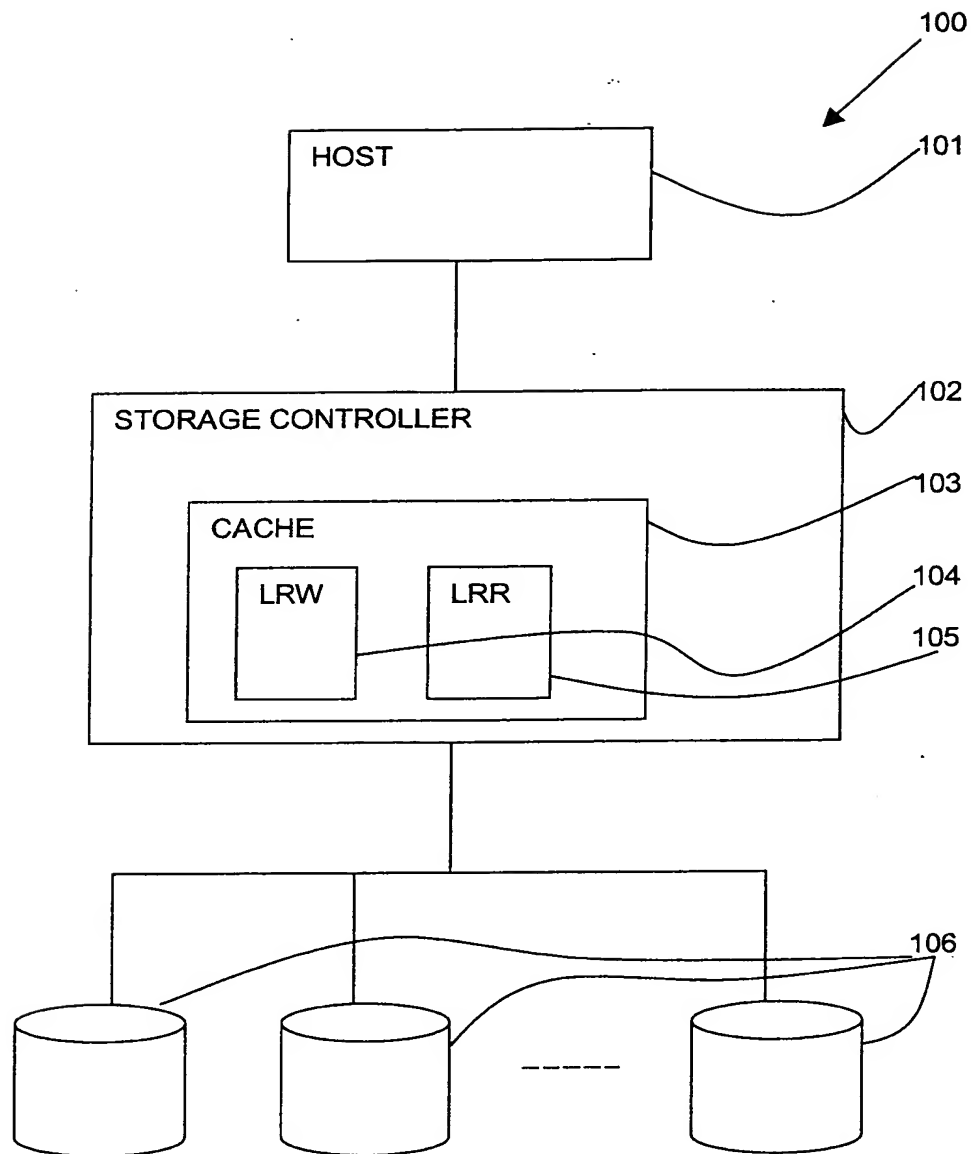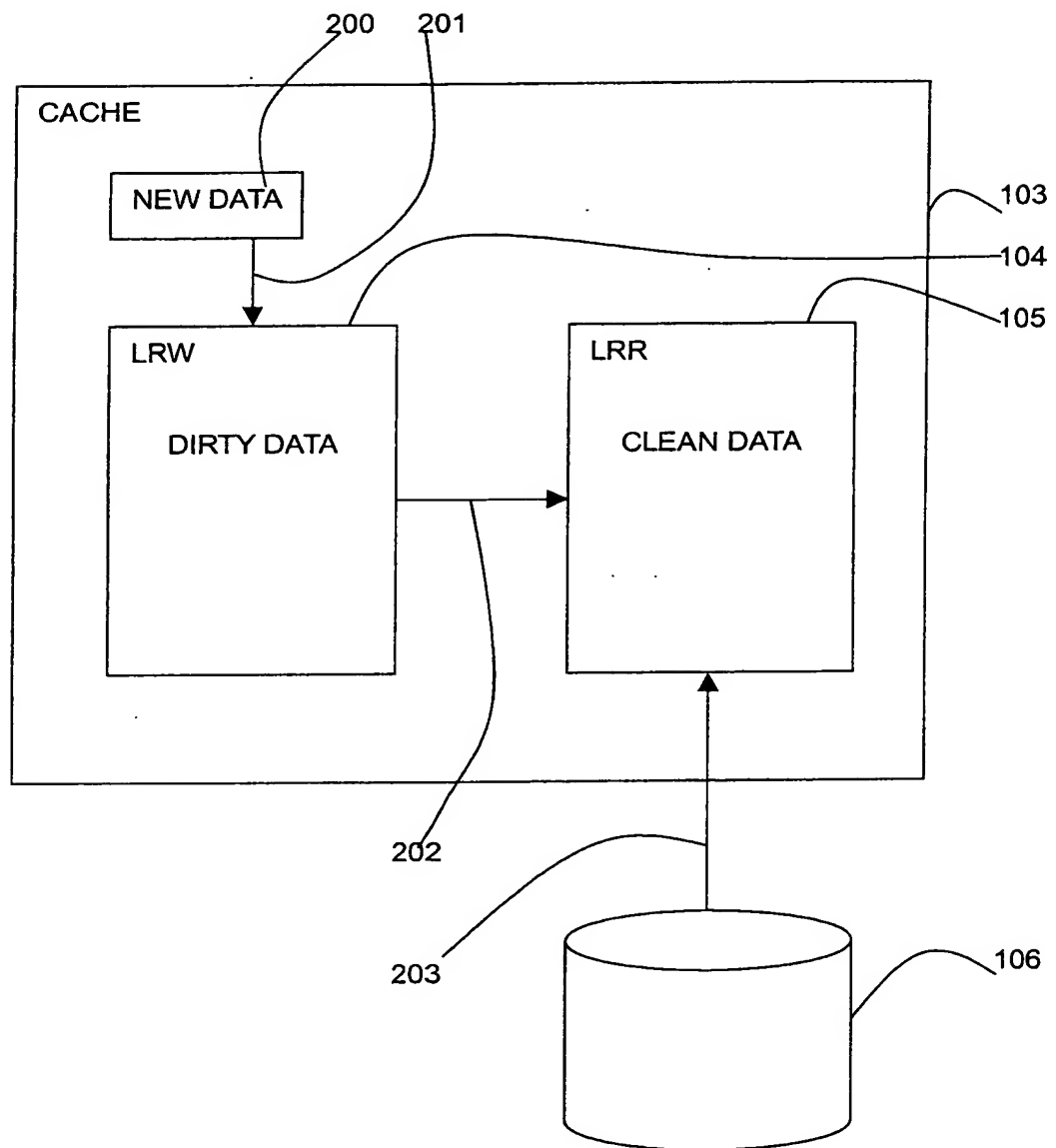
15

FIG. 1

FIG. 2

CACHE

200   201

NEW DATA

103

104

105

LRW

DIRTY DATA

LRR

CLEAN DATA

202

203

106

FIG. 3

DATA WRITE
RECEIVED IN CACHE — 301

↓

DATA INPUT AT HEAD
OF LRW LIST AS
DIRTY DATA — 302

↓

FLAG KEPT WITH DATA
DESCRIPTION INDICATING IF
DATA IS READ — 303

↓

DATA REACHES TAIL OF LRW
LIST AND IS DESTAGED TO
DISK — 304

↓

305

IS DATA
ALREADY IN
LRR LIST? — YES → LEAVE DATA
WHERE IT IS IN
LRR LIST — 306

NO

↓

307

HAS DATA BEEN
READ WHILST
DIRTY? — YES → SEND DATA TO
HEAD OF LRR
LIST — 308

NO

↓

DISCARD DATA — 309